Lab 5 Practice: Inference for a Single Mean

Overview

In this lab we use the one sample t-test to perform inference for a single mean. We also use the *t* distribution to (1) calculate p-values based on t test statistics, and (2) calculate t-scores used in confidence intervals for specific confidence levels.

The Data: Ethiopian adolescents

Dr. Craig Hadley, an anthropologist at Emory University, is the principle investigator of a study on the social determinants of wellbeing in adolescents in Ethiopia. This study is aimed at understanding youth social, cultural, educational and health trajectories: how these vary by place, what predicts different youth trajectories, and what are the determining factors that lead to trajectories with greater or less wellbeing. The data set provided contains basic demographic information, nutritional indicators, and economic indicators. The data set `ethiopia.csv` contains 9 variables.

`place`	city, rural area, or small town
`age`	age in years (13, 14, 15, 16, or 17)
`weight_kg`	weight in kilograms
`height_cm`	height in centimeters
`muac_cm`	mid-upper arm circumference (in centimeters)
`hhincome`	household income (in Ethiopian birr)
`gender`	male, female
`roof`	grass or iron/tin
`bmi`	body mass index (weight in kg divided by height in meters squared)

Practice

1. Open the `R Reference Guide` from the `Lab Content` area of PolyLearn.
2. Open `RStudio` and then open a brand new `Rmarkdown` document by clicking on the green plus sign on the top left of RStudio. Delete everything in the Rmarkdown document. Identify the `Rmarkdown` tab of your `R Reference Guide`. Copy and paste the three code chunks (header, set up chunk, and import data code chunk) into your Rmarkdown document.
3. Identify the `Lab Data Sets` folder on PolyLearn and download the `ethiopia` data set to a location on your computer (i.e., desktop, STAT 217 folder). Follow the steps in the `Importing` tab of the `R Reference Guide` to import the `ethiopia` data set and save your import code in an R chunk.
4. How many observations are in this data set?
5. What do you think is the target population for this study?
6. Examine the distribution of `bmi`. Be sure to comment on the shape, mean, and standard deviation.
7. Researchers want to know if the population of all Ethiopian adolescents have a healthy bmi or not, where a healthy bmi value is considered to be 18. Suppose you want to use a one sample t-test to determine if the population mean bmi differs from 18 on average, or use a confidence interval to estimate the population mean bmi. Are conditions satisfied for valid inference? Fully justify your answer. *Hint: There is no R code required to answer this question - do not actually execute the test yet!*
8. Perform the appropriate test to determine if the population mean bmi differs from 18 on average and use a 95% confidence interval to estimate the average bmi. Fill in the blanks in the following paragraph to summarize your results.

   When submitting answers for the paragraph below, please either:

- write just the answers that go in the blanks (do not copy whole paragraph), or
- copy the whole paragraph but mark your answers in bold using double asterisks, as in `**myanswer**`

In summarizing the data, we found that _______ *(insert #)* adolescents had an average bmi of _______ with a standard deviation of _______ . The distribution of bmi appeared to be (approximately normal/ right skewed/left skewed). We conducted a _______ test to determine if the population mean bmi differs from 18, and we evaluated the hypotheses H_0:_______ versus H_a:_______ . The conditions for valid inference (were/were not) satisfied. The test yielded a test statistic of $t =$_______ and a p-value of _______ . At the 0.05 level of signifcance, we (reject/fail to reject) the null hypothesis. We (do/do not) have evidence that the population _______ bmi differs from 18. We are 95% confident that the population mean bmi is in the interval _______ . Since this interval (does/does not) contain 18, it (is/is not) plausible that the population mean bmi is 18, indicating that the population mean of Ethiopian adolescents (less than/greater than/equal to) the healthy value of 18.

For the next two problems, it may be helpful to watch this brief video (2:38) demonstrating how to use the applet. https://youtu.be/Rwk5e9yTR7g

9. *Use the t-distribution to calculate p-values.* Suppose we were performing a hypothesis test with a **two-sided** alternative hypothesis. For this exercise, use the applet http://www.lock5stat.com/StatKey/theoretical_distribution/theoretical_distribution.html#t. For each numbered item below, **calculate** p-values for the given test statistic values and degrees of freedom and **state** if you would reject or fail to reject H_0 at the $\alpha = 0.05$ level. Lastly, circle or write in answers in the paragraph below. 2

   (a) $t = 1, df = 24$ _______ *reject / fail to reject* H_0

   (b) $t = 2, df = 24$ _______ *reject / fail to reject* H_0

   (c) $t = 3, df = 24$ _______ *reject / fail to reject* H_0

   (d) $t = 1, df = 99$ _______ *reject / fail to reject* H_0

   (e) $t = 2, df = 99$ _______ *reject / fail to reject* H_0

   (f) $t = 3, df = 99$ _______ *reject / fail to reject* H_0

   As the test statistic increases, the p-value (increases/decreases). Therefore, larger test statistics present (more/less) evidence against the null hypothesis. As the degrees of freedom increase, the p-value (increases/decreases). This is because as the degrees of freedom increase, the t distribution gets closer to a _______ distribution. The same test statistic value with different degrees of freedom (can/cannot) result in a different conclusion for a specified level of significance (e.g., $\alpha = 0.05$).

10. *Use the t-distribution to calculate t-scores for a confidence interval.* Use the applet http://www.lock5stat.com/StatKey/theoretical_distribution/theoretical_distribution.html#t to calculate a t-score to create 90, 95, and 99% confidence intervals with 24 and 99 degrees of freedom. Please report the **positive** version of the t-score. Circle or write in answers in the paragraph below. 2

    (a) 90%, $df = 24$ _______

(b) 95%, $df = 24$ ________

(c) 99%, $df = 24$ ________

(d) 90%, $df = 99$ ________

(e) 95%, $df = 99$ ________

(f) 99%, $df = 99$ ________

As the confidence level increases, the t-score (increases/decreases). This means that higher confidence levels result in (narrower/wider) confidence intervals. As the degrees of freedom increase, the t-score (increases/decreases). This means that for the same confidence level, larger degrees of freedom result in (narrower/wider) confidence intervals.

11. *Challenge question: This question is optional and you are not required to submit an answer to this question.* Perform a one-sample t-test on the `bmi` variable testing H_0: $\mu = 17.9$ versus H_a: $\mu \neq 17.9$ using $\alpha = 0.05$ (so all we are doing is changing the value tested in the null hypothesis). Compare your R output to the output from the Question 8. Did changing the value of μ_0 affect (i) the test statistic, and (ii) the p-value? If any changes are observed, explain why.

12. Submit this lab assignment as an html compiled from R Markdown. Make sure all names of group members who contributed to this lab assignment are on the html file.